þÿ F i S T t o w a r d s a F r e e S e m a n t i c T a g g e r o f M o d e

Kettunen, Kimmo Tapio

The Association for Computational Linguistics
2019-01-30

þÿ K e t t u n e n , K T 2 0 1 9 , F i S T t o w a r d s a F r e e S e m a n t i c T a g g e r o f M o d e r
in The fifth International Workshop on Computational Linguistics for Uralic Languages
Proceedings of the Workshop . The Association for Computational Linguistics , Stroudsburg ,
pp. 66-76 , International Workshop on Computational Linguistics for Uralic Languages ,
Tartu , Estonia , 07/01/2019 . https://doi.org/10.18653/v1/w19-0306

http://hdl.handle.net/10138/306801
https://doi.org/10.18653/v1/w19-0306

FiST – towards a Free Semantic Tagger of Modern Standard Finnish[1]

Kimmo Kettunen [0000-0003-2747-1382]
The National Library of Finland, DH Research
firstname.lastname@helsinki.fi

Abstract

This paper introduces a work in progress for implementing a free full text semantic tagger for Finnish, FiST. The tagger is based on a 46 226 lexeme semantic lexicon of Finnish that was published in 2016. The basis of the semantic lexicon was developed in the early 2000s in an EU funded project Benedict (Löfberg et al., 2005). Löfberg (2017) describes compilation of the lexicon and evaluates a proprietary version of the Finnish Semantic Tagger, the FST[2]. The FST and its lexicon were developed using the English Semantic Tagger (The EST) of University of Lancaster as a model. This semantic tagger was developed at the University Centre for Corpus Research on Language (UCREL) at Lancaster University as part of the UCREL Semantic Analysis System (USAS[3]) framework. The semantic lexicon of the USAS framework is based on the modified and enriched categories of the *Longman Lexicon of Contemporary English* (McArthur, 1981).

We have implemented a basic working version of a new full text semantic tagger for Finnish based on freely available components. The implementation uses Omorfi and FinnPos for morphological analysis of Finnish words. After the morphological recognition phase words from the 46K semantic lexicon are matched against the morphologically unambiguous base forms. In our comprehensive tests the lexical tagging coverage of the current implementation is around 82–90% with different text types. The present version needs still some enhancements, at least processing of semantic ambiguity of words and analysis of compounds, and perhaps also treatment of multiword expressions. Also a semantically marked ground truth evaluation collection should be established for evaluation of the tagger.

Tiivistelmä

Suomessa on harjoitettu kieliteknologiaa laaja-alaisesti 1980-luvun alusta, ja melkein 40 vuotta jatkunut tutkimus ja kehitystyö on tuottanut useita merkittäviä ohjelmistoja suomen kielen analyysiin. Alkuvuosikymmenien käytöltään rajoitetuista ohjelmistoista on siirrytty 2000-luvulla paljolti joko avoimen lähdekoodin ohjelmiin tai ohjelmien vapaaseen saatavuuteen. Vapaasti saatavia suomen kielen keskeisiä kieliteknologisia ohjelmia on olemassa tällä hetkellä hyvin morfologiseen ja syntaktiseen analyysiin, esimerkiksi *Omorfi*, *Voikko ja FinnPos* morfologiaan ja *Finnish depedency parser* lauseenjäsennykseen. FiNER-ohjelmistolla voidaan tunnistaa ja merkitä erisnimiä. Toistaiseksi ei kuitenkaan ole olemassa ainoatakaan vapaasti saatavaa suomenkielisten kokotekstien kattavaa semanttista merkintää tekevää ohjelmaa, semanttista taggeria. Voikin todeta, että suomen kielen automaattiseen semanttiseen käsittelyyn on jäänyt jos ei aivan tyhjiö, niin kuitenkin suuri aukko.

Tässä julkaisussa esitellään FiST, työn alla oleva suomen nykykielen kokotekstien semanttinen merkitsin. FiSTin ensimmäinen versio perustuu vapaasti saatavilla oleviin osiin: 46 226 sanan semanttiseen leksikkoon sekä vapaisiin morfologisen analyysin ohjelmiin Omorfiin ja FinnPosin. Ohjelma merkitsee teksteihin sanojen semanttisia luokkia noin 82–90 %:n sanastollisella kattavuudella.

[2] The tagger was implemented by Kielikone Ltd. It used proprietary analysis modules of Kielikone for morphological and morpho-syntactic analysis of Finnish. The software has not been publicly available and can be considered partly outdated now. Its operational design is described in Löfberg et al. (2005).

[3] http://ucrel.lancs.ac.uk/usas/

Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages, pages 66–76,
Tartu, Estonia, January 7 - January 8, 2019.

1 Introduction

Language technological resources for analysis of written modern standard Finnish can be considered reasonably good overall. However, a major aspect of automatic analysis of written Finnish is still poorly covered as there is no freely available full text semantic analyzer or tagger of Finnish. Lack of semantic resources for Finnish was already noted in the META NET white paper (Koskenniemi et al., 2012). The situation has not improved noticeably since the publication of the META NET report, although some semantic lexical resources have been published in recent years.

Computational linguistics has been practiced in Finland since the early 1980s. During the last almost four decades several important analysis software for Finnish morphology and syntax have been produced. Without going too deeply in to historical details, early implementations include e.g. the first full computational morphological model for Finnish, TWOL (Koskenniemi, 1983), and a general syntactic parsing formalism Constraint Grammar (CG, Karlsson, 1990). In the 21st century most of the major new linguistic analysis tools have become either open source or at least freely available or usable. Such programs are, e.g., free morphological analyzers Omorfi[4] (Pirinen, 2015), Voikko[5] and FinnPos[6] (Silfverberg et al., 2016). A free dependency parser for Finnish is provided by the BioNLP group at the University of Turku (Turku Neural Parser Pipeline[7]). The Language Bank of Finland[8] provides also access to these and other tools, such as a Finnish named entity tagger FiNER.

As good as these tools may be in their tasks, they serve only a quite limited function. Morphological and syntactic analyses are rarely goals in themselves in real life text analysis outside linguistics; morphological and syntactic analyses serve only as mid-level results for further processing of textual content. In information oriented parlance, these tools do not reveal anything about the content of the texts or their aboutness. Most of the time contents of the texts are interesting for research outside linguistics, not the linguistic form. Proper content analysis tools for Finnish are scarce. Out of the existing tools only FiNER has limited semantic capabilities, as it marks names and name like entities.

A few semantically oriented lexicons have also been compiled and published for Finnish, namely FinnWordnet (Lindén and Carlson, 2010; Lindén and Niemi, 2016) and FrameNet (Lindén et al., 2017). Some type of semantic analyzers for Finnish could be produced using FinnWordnet, but so far usage of FinnWordnet for semantic level analyses seems to have been non-existent. WordNets are also not comprehensive semantic lexicons for full text analysis: they contain only words belonging to four main word classes, nouns, verbs, adjectives and adverbs. Their contents seem also a bit problematic. FrameNet, on the other hand, is an even more restricted description of a set of situations, entities and relationships of the participants (lexemes) in a lexical frame. A third available lexical tool, YSO[9], General Finnish Ontology, serves mainly indexing of Finnish cultural content (Hyvönen et al., 2008). Ontologies are useful for many purposes, but they are mainly non-linguistic descriptions (Hirst, 2004) and do not even aim to cover all word classes of natural language. Thus they do not suit for full text analysis. A proper semantically oriented full text analyzer of Finnish would improve possibilities of textual content analysis vastly.[10]

[4] https://github.com/flammie/omorfi
[5] https://voikko.puimula.org/
[6] https://github.com/mpsilfve/FinnPos
[7] https://github.com/TurkuNLP/Turku-neural-parser-pipeline
[8] https://www.kielipankki.fi/language-bank/
[9] http://finto.fi/yso/en/?clang=fi
[10] A few other approaches can also be mentioned here. Besides YSO several smaller subject matter ontologies exist, e.g. AFO (agriculture and forestry), JUHO (Government) etc. (https://seco.cs.aalto.fi/ontologies/). Haverinen (2014) introduces semantic role labeling for Finnish. This is related to argument structure of verbs in syntactic parsing of sentences and is of limited semantic value. BabelNet (https://babelnet.org/, Navigli and Ponzetto, 2012) is a large multilingual encyclopedic database, which includes also Finnish. Its descriptions for words have been collected from multilingual Wikipedia articles using WordNet.

This paper introduces a work in progress for implementing a free full text semantic tagger for Finnish, FiST. The tagger is based on freely available morphological processors and a 46 226 lexeme semantic lexicon of Finnish that was published in 2016. We shall first discuss semantic tagging in general and design of FiST. After that we evaluate lexical coverage of the tagger with different types of available digital Finnish corpora. Finally, we discuss improvements needed for the tagger and conclude the paper.

2 Semantic Tagging

Semantic tagging is defined here as a process of identifying and labelling the meaning of words in a given text according to some semantic scheme. This process is also called semantic annotation, and in our case it uses a semantic lexicon to add labels or tags to the words. (Leech, 2003; Löfberg, 2017; Wilson and Thomas, 1997).

Semantic tagging discussed here is based on the idea of semantic (lexical) fields. Wilson and Thomas (1997, p. 54) define a semantic field as "a theoretical construct which groups together words that are related by virtue of their being connected – at some level of generality – with the same mental concept". In other words "a semantic field is a group of words which are united according to a common basic semantic component" (Dullieva, 2017, formulating Trier's insight of semantic fields; cf. also Lutzeier, 2006; Geeraerts, 2010). Semantic lexicon of USAS is divided in to 232 meaning classes or categories which belong to 21 upper level fields. Figure 1 shows one upper level semantic field, *Money & Commerce*, and its meaning classes (USAS Semantic Tag Set for Finnish). Alphanumeric abbreviations in front of the meaning classes are the actual hierarchical semantic tags used in the lexicon. According to Piao et al. (2005), the depth of the semantic hierarchical structure is limited to a maximum of three layers, since this has been found to be the most feasible approach.

I	MONEY & COMMERCE
I1	Money generally
I1.1	Money: Affluence
I1.2	Money: Debts
I1.3	Money: Price
I2	Business
I2.1	Business: Generally
I2.2	Business: Selling
I3	Work and employment
I3.1	Work and employment: Generally
I3.2	Work and employment: Professionalism
I4	Industry

Figure 1. Semantic field of Money & Commerce in the USAS Finnish semantic lexicon

The major 21 discourse fields used in the USAS are shown in Figure 2[11].

[11] http://ucrel.lancs.ac.uk/usas/

A	General & Abstract Terms
B	The Body & the Individual
C	Arts & Crafts
E	Emotional Actions, States & Processes
F	Food & Farming
G	Government & the Public Domain
H	Architecture, Building, Houses & the Home
I	Money & Commerce
K	Entertainment, Sports & Games
L	Life & Living Things
M	Movement, Location, Travel & Transport
N	Numbers & Measurement
O	Substances, Materials, Objects & Equipment
P	Education
Q	Linguistic Actions, States & Processes
S	Social Actions, States & Processes
T	Time
W	The World & Our Environment
X	Psychological Actions, States & Processes
Y	Science & Technology
Z	Names & Grammatical Words

Figure 2. Top level domains of the USAS tag set

This top level domain and its subdivisions were developed from the categories used in the Longman Lexicon of Contemporary English (LLOCE, McArthur, 1981). LLOCE uses 14 top level domains. Some of those were divided into more fine-grained classes in the USAS. Also one more class, *Names and Grammatical words*, was added (Archer et al., 2004).

3 The Finnish Semantic Lexicon

The core of this kind of approach to semantic tagging is naturally the semantically marked lexicon. Semantic lexicons using the USAS schema have so far been published in 12 languages (Multilingual USAS; Piao, 2016[12]). Out of these lexicons the Finnish lexicon is the most comprehensive and mature. It has been compiled manually, as many of the lexicons for other languages are compiled partly or wholly automatically based on the USAS English lexicon and bilingual dictionaries. In different evaluations the Finnish lexicon has been shown to be capable of dealing with most general domains which appear in modern standard Finnish texts (Löfberg, 2017; Piao et al., 2016). Furthermore, although the semantic lexical resources were originally developed for the analysis of general modern standard Finnish, evaluation results have shown that the lexical resources are also applicable to analysis of both older Finnish texts and the more informal type of writing found on the Web. The semantic lexical resources can also be tailored for various domain-specific tasks thanks to the flexible USAS category system. Lexemes can be added to the lexicon easily, as it is open and its description is fairly straightforward.

The Finnish semantic lexicon consists of 46 226 lexemes. Out of these about 58% are nouns, 7% verbs, 17% proper names, 7% adjectives and 7% adverbs (Löfberg, 2017).[13] Rest of the words belong to small fixed classes. Löfberg (2017: Table 7, 139) lists the distribution of lexical entries

[12] The list of the 11 other languages is: Arabic, Chinese, Czech, Dutch, Italian, Malay, Portuguese, Russian, Spanish, Urdu and Welsh. Sizes of the lexicons vary between 1 800 and 64 800 single word entries. Finnish lexicon is thus the third largest of all available after lexicons of Malay and Chinese (Piao et al., 2016). Eight of the languages have an existing semantic tagger. Those that do not have are Arabic, Malay, Urdu and Welsh.

[13] Distributions for POS categories are given in Löfberg (2017, Table 4, 135). The size of the lexicon in the thesis is slightly smaller than the size of the published lexicon.

in the top level semantic categories in the single word lexicon of the FST. The table is too large to be shown here, so we list only the five categories that have most lexemes. The largest category is Z (Names and Grammatical Words), with 9 755 lexical entries (21.31%). Second largest category is A (General & Abstract Terms) with 4 544 entries (9.93%). The third largest category is B (The Body & The Individual) with 3 734 entries (8.16%). A (Social Actions, States & Processes) and L (Life & Living Things) are the next ones with 3 401 (7.43) and 2 798 (6.11%) entries, respectively. These five categories constitute about 52 per cent of the entries in the lexicon.

4 Design of FiST

Our current implementation of FiST is simple and straightforward. It uses existing free morphological tools, Omorfi and FinnPos, for morphological analysis and disambiguation of input texts. After the morphological phase words of the input text are unambiguous and in base form, and the tagger tries to match the words to lexical entries in its semantic lexicon. If a word is found in the lexicon, it is tagged and returned with word class and the semantic label(s) found. If the word is not in the semantic lexicon, it is marked as Z99, unknown, and returned with this tag and the morphological analysis for the word, if such is available. Figure 3. shows the working process of FiST.

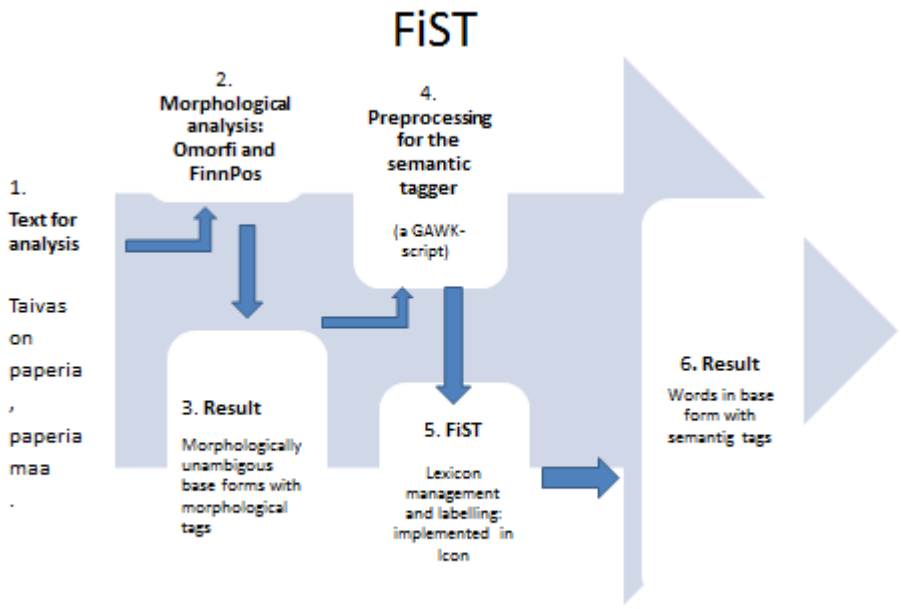

Figure 3. FiST schematically

The analysis result of the opening verse of the first poem of Pentti Saarikoski's first poetry collection in word per line form is shown in Table 1.

Input	Output of FiST	Explanation
Taivas	taivas Noun W1 S9 Z4	A noun with three semantic tags: the first one is the right one (*The Universe*).
on	olla Verb A3+ A1.1.1 M6 Z5	A verb with four semantic tags: the first one denoting to existence is the right one.
paperia	paperi Noun O1.1 Q1.2 B4 P1/Q1.2	A noun with four semantic tags: the first one denoting to solid matter, O1.1, would be the best choice.
,	PUNCT	Punctuation
paperia	paperi Noun O1.1 Q1.2 B4 P1/Q1.2	A noun with four semantic tags: the first one denoting to solid matter, O1.1, would be the best choice.
maa	Maa Noun M7	An unambiguous noun denoting to areas.
.	PUNCT	Punctuation

Table 1. FiST's analysis of the first verse of a poem by Pentti Saarikoski

5 Evaluation

As there is no semantically marked evaluation collection available, we have not been able to evaluate FiST's semantic accuracy so far. However, we have performed quite thorough testing of the current implementation's lexical coverage. Our evaluation data consists of 17 texts that range from about 42 000 words to ca. 28.6 million words, the largest corpus being the Finnish part of the Europarl corpus v6[14]. We show also morphological recognition rates for all except one of the texts with Omorfi. This gives an idea of the coverage of the semantic lexicon in comparison to the lexicon of a morphological analyzer, which is usually much larger. Omorfi's lexicon is almost ten times larger than the semantic lexicon – it consists of 424 259 words (Pirinen, 2015). Our formula for coverage of FiST is the following: *(100* (1-(missed tag/(NR-comma-number))))*. Here missed tags are those words that are tagged as Z99, unknown. Punctuation marks and numbers are subtracted from the number of records/words (NR). Input for the evaluation is one tagged word/line, with no empty lines.

Figures 4 and 5 show tagging results of our current tagger version with 17 texts. Figure 4 shows results of modern texts, and Figure 5 results of older texts.

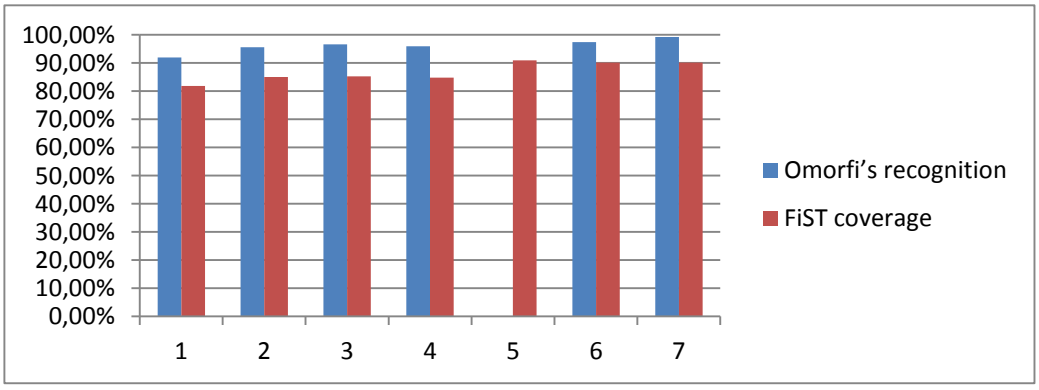

Figure 4. Coverage of semantic tagging of FiST with different modern Finnish texts (N.B. morphological recognition rate for Europarl is not available)

[14] http://www.statmt.org/europarl/archives.html

In Figure 4 text #1 is Suomi24[15] discussion forum data (494 000 tokens), texts #2-4 are sentences from news in the Leipzig corpus[16] (100K, 300K and 1M tokens), text #5 is Europarl v.6 text (ca. 28.6 M tokens), #6 prose of Pentti Saarikoski (172 920 tokens, not publicly available) and text #7 is sample sentences of FinnTreebank[17] (examples from a Finnish grammar, 138 949 tokens).

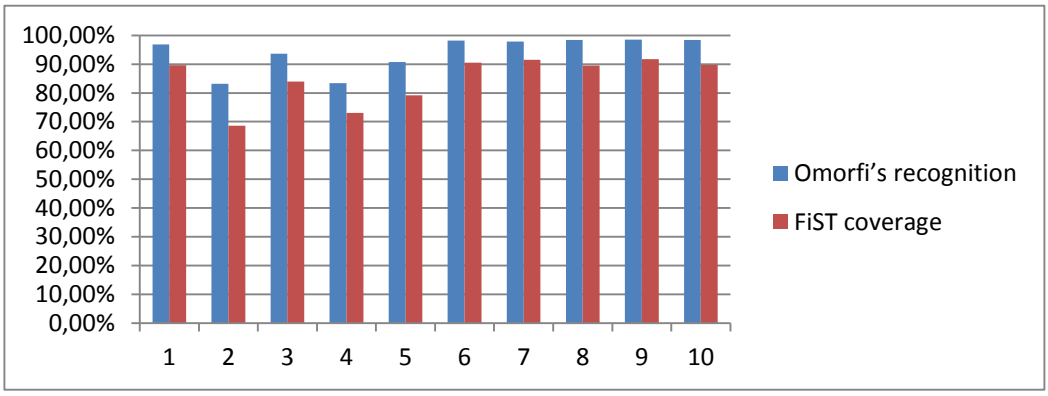

Figure 5. Coverage of semantic tagging of FiST with different older Finnish texts

In Figure 5 text number 1 is Bible translation of 1938[18] (544 474 tokens), #2 is newspaper/journal Turun Wiikkosanomat 1831 (60 390 tokens), #3 newspaper/journal Mehiläinen 1859 (154 370 tokens), #4 newspaper/journal Oulun Viikko-Sanomia 1841 (68 491 tokens), and #5 is newspaper/journal Sanansaattaja Wiipurista 1841 (49 802 tokens). All the journalistic texts are from digital collection of the Institute for the Languages of Finland [19]. Texts #6–#10[20] are literary works of Finnish authors Juhani Aho, Minna Canth, Arvid Järnefelt, Teuvo Pakkala, and Kyösti Wilkuna from late 19th and early 20th century with 42 000–334 000 tokens. They are also from the collection of the Institute for the Languages of Finland. These collections are manually edited.

Results of the analyses show that FiST is capable of annotating texts of modern standard Finnish quite well already now. With many of the texts about 90% of the words get a semantic label in FiST's analysis. This applies also to literary texts of Pentti Saarikoski, both prose and poetry. Proceedings of the European Parliament v6 (Koehn, 2005), our largest data collection, gets also a high coverage: 90.9%. Suomi24 data and data from the Leipzig corpus obtain clearly lower coverage. This is mainly due to the nature of the texts. Suomi24 contains informal discussions that may include lots of misspelled words, slang and foreign words. Texts of the Leipzig corpus have been crawled from the Web automatically and may thus contain more noise, i.e. misspellings, control characters, HTML code etc. (Quasthoff et al., 2006).

Older literary texts and the Bible translation of 1938, however, obtain a quite good coverage, round 90%, as can be seen in Figure 5. Our oldest texts are from 1831–1859, four newspapers: Turun Wiikko-Sanomia (1831), Oulun Viikko-Sanomia (1841), Sanansaattaja Wiipurista (1841) and Mehiläinen (1859). These versions are manually edited clean versions from the Institute for the Languages of Finland. Considering the age of the data, these get also quite good coverage with FiST, 68.6, 73, 79.22 and 84 per cent.

[15] http://metashare.csc.fi/repository/browse/the-suomi-24-corpus-2015h1/b4db73da85ce11e4912c005056be118ea699d93902fa49d69b0f4d1e692dd5f1/
[16] http://wortschatz.uni-leipzig.de/de/download
[17] http://www.ling.helsinki.fi/kieliteknologia/tutkimus/treebank/sources/
[18] http://raamattu.fi/1933,38/
[19] http://kaino.kotus.fi/korpus/1800/meta/1800_coll_rdf.xml
[20] http://kaino.kotus.fi/korpus/klassikot/meta/klassikot_coll_rdf.xml

We performed also a few small scale test runs with our text data using one available lexicon to get more insight into lexical coverage. The so called Kotus wordlist[21] which contains ca. 94 000 lexemes from a dictionary of modern Finnish, has a good coverage: it was only a few per cent units below coverage of the semantic lexicon. When we tested coverage by combining the semantic lexicon and words of the Kotus wordlist not included in the semantic lexicon, we noticed an increase of few per cent units in matching of our evaluation data. Our tests were performed with three small texts, and the tests are thus not as comprehensive as our tests with FiST's main version, but clearly indicative.

Lindén and Niemi (2016) have evaluated FinnWordnet's lexical coverage with samples. In a large newspaper corpus (of unspecified size) coverage was 57.3%. If only nouns, verbs, adjectives and adverbs were counted and proper names excluded, the coverage was 82.4%. Thus FinnWordnet's lexicon is probably not sufficient for good lexical coverage of Finnish texts as an only lexical resource. On the other hand, lexical coverage of the semantic lexicon of FiST could be increased with a few per cent units by adding lexemes to it from other available lexicons. This, of course, would also mean laborious semantic marking of the additions.

6 Discussion

The current implementation of FiST is a simplified basic version of a semantic tagger. It lacks at least two main components: word sense disambiguation (WSD) and proper handling of compounds. The semantic lexicon of Finnish marks ambiguous meanings of words by giving several meaning tags. Word *huone* ('room'), for example, is given an entry *huone Noun H2 S9*. Parts of buildings belong to class H2, and S9 is for words that have a religious or supernatural meaning. The primary meaning of *huone* is H2, but in some contexts, especially in astrology, it could be S9. Thus semantic disambiguation would be needed to be able to distinguish meanings of ambiguous words. Word sense disambiguation has gained lots of interest in computational linguistics during the past 20 years, and thus ways to disambiguate ambiguous words should be found with a reasonable effort (Edmonds, 2006). Rayson et al. (2004), e.g., describe several methods they use for WSD in the English Semantic Tagger. A few most simple ones of these should be easy to implement.

If the word is ambiguous, i.e. it has more than one sense, the different senses are listed in the lexicon arranged in perceived frequency order (Löfberg, 2017: 74). The earlier example from analysis of the poem of Pentti Saarikoski shows this: *paperi Noun O1.1 Q1.2 B4 P1/Q1.2*. In the analysis we can also see an example of so called "slash tag" (or "portmanteau tag") of the USAS framework. The slash shows that the word belongs to more than one category. *Paperi* belongs to solid matter, but also to category of education (P1) and literary documents and writing, Q1.2. A counting in the lexicon shows that 7 791 lexemes have been described as ambiguous and 10 556 have the slash tag. Out of the ambiguous lexemes 5 476 have two meanings, and 1 449 three meanings. There are almost 500 words with four meanings and almost 200 with five, but after six meanings number of lexemes having more meanings drops to tens. The more meanings the lexeme has been given, the more abstract it tends to be. Abstract nouns like *meneminen* ('going') and *tuleminen* ('coming') have 10 meanings in the lexicon. 85% of the slash category words have only one slash tag.

Another deficiency in the FiST's implementation is handling of compounds. Finnish is notoriously rich in compounds, and no lexicon can contain all of them. The Finnish semantic lexicon includes the most common compounds as such, but for those that are not included, the meaning should be composed out of the meanings of component parts. *Kivitalo* ('house made of stone/concrete'), for example, is not in the lexicon, but its parts are. *Kivi* is ***Noun O1.1 B5***, and *talo* ***Noun H1 S5+ I3.1/M.*** In practice the semantic marking should be *kivitalo* ***H1/O1.1***, as the most meaningful part of the compound is usually the last part. For this to succeed, much depends on the morphological analyzer, as it analyzes and splits the compounds for the semantic tagger.

[21] http://kaino.kotus.fi/sanat/nykysuomi/

It would be desirable, that the morphological analyzer returned compounds both as wholes and split, as it would make search of available compounds in the semantic lexicon easier.[22]

A third possible improvement for the FiST would be handling of multiword expressions (MWEs) that consist of two or more separate orthographic words. *Englannin kanaali*, *Euroopan Unioni* and *musta pörssi* are some examples of MWEs. The original FST (and the EST) has a separate lexicon of over 6000 entries for multi word expressions (Löfberg, 2017).This lexicon has not been published. A list of MWEs could be compiled with a reasonable effort, but semantic description of thousands of words would take time, especially as a substantial part of the MWEs are non-compositional idiomatic expressions (Piao et al., 2016). The Finnish Wordnet, for example, contains about 13 000 multiword nouns, but considering that the lexicon was produced as a direct translation of the English Wordnet, many of the MWES do not seem very frequent or crucial to Finnish.

7 Conclusion

We have described in this paper FiST, a first version of a full text semantic tagger for Finnish. We have provided background for the tagger's lexical semantic approach and evaluated its capabilities mainly as a semantic tagger of modern standard Finnish. The first results can be considered promising: the tagger works robustly even with large data of millions of words and achieves a good lexical coverage with many types of texts. Our evaluation of FiST confirms also that the Finnish semantic lexicon of USAS is of high quality and it covers also data from time periods that are supposedly out of its main scope.

Semantic tagging can be used in many natural language processing applications, such as terminology extraction, machine translation, bilingual and multilingual extraction of multi-word expressions, monolingual and cross-lingual information extraction, as well as in automatic generation, interpretation, and classification of language. Semantic tagging with the English Semantic Tagger of UCREL has been successfully utilized for content analysis, analysis of online language, training chatbots, ontology learning, corpus stylistics, discourse analysis, phraseology, analysis of interview transcripts, and key domain analysis (Löfberg, 2017; http://ucrel.lancs.ac.uk/usas/; http://ucrel.lancs.ac.uk/wmatrix/#apps). These kinds of applications could also be targets for FiST.

In the future we wish to improve the tagger's performance with the still missing features. If possible, we evaluate the tagger's semantic accuracy with semantically marked data. We also believe that even the current plain implementation is suitable for many textual analysis purposes, e.g. content wise topic analysis (vs. statistical, where words are only strings without meaning), lexical content surveying, semantically oriented lexical statistics etc. We have also performed some trials to use data tagged with FiST as training data for a machine learning algorithm to learn a semantic tagger for Finnish. So far our trials have not been very successful due to the rich feature set of semantic marking. Most of the standard machine learning environments we have tried run out of memory with the number of features of semantically tagged data. Probably at least some smaller scale niche semantic field analyzer could be developed with marked data provided by FiST.

Acknowledgements

We wish to thank Dr. Laura Löfberg for useful comments and providing some of her evaluation data for use. Our largest data file, Europarl v6, was analyzed in Taito cluster of the CSC - It Center For Science Ltd.

[22] We have been using FinnPos from the Mylly resources (https://www.kielipankki.fi/support/mylly/) of the Language Bank of Finland. Currently FinnPos does not split compounds to their parts, although it analyzes the base forms of the wholes. Omorfi splits compounds, as does Voikko, too. Voikko's splitting, however, does not seem very useful, as it separates also sub-word parts, e.g. derivational endings, in the output. Omorfi and Voikko do not disambiguate the different morphological interpretations, which makes usage of either of them as sole morphological component complicated.

References

Dawn Archer, Paul Rayson, Scott Piao, Tony McEnery. 2004. Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. In Williams G. and Vessier S. (eds.) Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004), Lorient, France, 6-10 July 2004. Université de Bretagne Sud. Volume III, pp. 817-827. ISBN 2-9522-4570-3.

Karina Dullieva. 2017. Semantic Fields: Formal Modelling and Interlanguage Comparison. Journal of Quantitative Linguistics, 24:1, 1-15. DOI: 10.1080/09296174.2016.1239400

Philip Edmonds. 2006. Disambiguation. In Allan, K. (ed.), Concise Encyclopedia of Semantics, 223–239. Elsevier.

Dirk Geeraerts. 2010. Theories of Lexical Semantics. Oxford: Oxford University Press.

Katri Haverinen. 2014. Natural Language Processing Resources for Finnish Corpus Development in the General and Clinical Domains. TUCS Dissertations No 179. https://www.utupub.fi/bitstream/handle/10024/98608/TUCSD179Dissertation.pdf?sequence=2&isAllowed=y

Grame Hirst. 2004. Ontology and the lexicon. In Staab S., Studer R. (eds.) Handbook on Ontologies. International Handbooks on Information Systems. Springer, Berlin, Heidelberg

Eero Hyvönen, Kim Viljanen, Jouni Tuominen, Katri Seppälä. 2008. Building a National Semantic Web Ontology and Ontology Service Infrastructure –The FinnONTO Approach. In: Bechhofer S., Hauswirth M., Hoffmann J., Koubarakis M. (eds) The Semantic Web: Research and Applications. ESWC 2008. Lecture Notes in Computer Science, vol 5021. Springer, Berlin, Heidelberg

Fred Karlsson. 1990. Constraint Grammar as a Framework for Parsing Unrestricted Text. H. Karlgren, ed., Proceedings of the 13th International Conference of Computational Linguistics, Vol. 3. Helsinki 1990, 168–173.

Philip Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit 2005.

Kimmo Koskenniemi. 1983. Two-level Morphology: A computational model for wordform recognition and production. Publications of the Department of General Linguistics, University of Helsinki 11. Helsinki: University of Helsinki.

Kimmo Koskenniemi et al. 2012. The Finnish Language in the Digital Age. META NET White paper series. http://www.meta-net.eu/whitepapers/e-book/finnish.pdf/view?searchterm=Finnish

Geoffrey Leech. 2004. Developing Linguistic Corpora: a Guide to Good Practice Adding Linguistic Annotation. https://ota.ox.ac.uk/documents/creating/dlc/chapter2.htm

Krister Lindén, Lauri Carlson. 2010. FinnWordNet – WordNet på finska via översättning. LexicoNordica – Nordic Journal of Lexicography, 17:119–140.

Krister Lindén, Jyrki Niemi. 2014. Is it possible to create a very large wordnet in 100 days? An evaluation. Language Resources and Evaluation, 48(2), 191–201.

Krister Lindén, Heidi Haltia, Juha Luukkonen, Antti O Laine, Henri Roivainen, Niina Väisänen. 2017. FinnFN 1.0: The Finnish frame semantic database. Nordic Journal of Linguistics, 40(3), 287-311.

Peter R Lutzeier. 2006. Lexical fields. In Allan, K. (ed.), Concise Encyclopedia of Semantics, 470–473. Elsevier.

Laura Löfberg, Scott Piao, Paul Rayson, Jukka-Pekka Juntunen, Asko Nykänen, Krista Varantola. 2005. A semantic tagger for the Finnish language. http://eprints.lancs.ac.uk/12685/1/cl2005_fst.pdf

Laura Löfberg. 2017. Creating large semantic lexical resources for the Finnish language. Lancaster University, 2017. 422 pages. http://www.research.lancs.ac.uk/portal/en/publications/creating-large-semantic-lexical-resources-for-the-finnish-language(cc08322c-f6a4-4c2b-8c43-e447f3d1201a)/export.html

Tom McArthur. 1981. Longman Lexicon of Contemporary English. Longman, London.

Multilingual USAS. https://github.com/UCREL/Multilingual-USAS

Roberto Navigli, Simone Paolo Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193:217–250.

Scott Piao, Dawn Archer, Olga Mudraya, Paul Rayson, Roger Garside, Tom McEnery, Andrew Wilson. 2005. A Large Semantic Lexicon for Corpus Annotation. In proceedings of the Corpus Linguistics 2005 conference, July 14-17, Birmingham, UK. Proceedings from the Corpus Linguistics Conference Series on-line e-journal, Vol. 1, no. 1, ISSN 1747-9398

Scott Piao et al. 2016. Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. In Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC2016), Portoroz, Slovenia, 2614–2619.

Tommi A Pirinen. 2015. Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. SKY Journal of Linguistics, vol 28, 381–393. http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Pirinen.pdf

Uwe Quasthoff, Matthias Richter, Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006, Genoa, 1799–1802.

Paul Rayson, Dawn Archer, Scott Piao, Tom McEnery. 2004. The UCREL semantic analysis system. In Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.

Miikka Silfverberg, Teemu Ruokolainen, Krister Lindén, Mikko Kurimo. 2016. FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish. Lang Resources & Evaluation 50: 863–878. https://doi.org/10.1007/s10579-015-9326-3

USAS Semantic Tag Set for Finnish. https://github.com/UCREL/Multilingual-USAS/raw/master/Finnish/USASSemanticTagset-Finnish.pdf

Andrew Wilson, Jenny Thomas. 1997. Semantic annotation. In Garside, R., Leech, G., & McEnery, T. (Eds.), Corpus annotation: Linguistic information from computer text corpora (pp. 53–65). New York: Longman.